Introduction to Statistics

Midterm Exam
Spring 2010

Name____

Show your work!

1) Given the following set of numbers:

X = {7.1, 1.8, 2.1, 1.5, 2.1}

a) Find the median, mean and standard deviation. State the formulas (or procedure) for these values, plug the numbers into the formula, and evaluate the results. You may use your calculator's automated functions to verify your answers. (8 points)

Reorder: 1.5, 1.8, 2.1,2.1,7.1
Median = 2.1
Mean

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1.5 + 1.8 + 2.1 + 2.1 + 7.1}{5} = 2.9$$

Standard deviation

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(1.5-2.9)^2 + (1.8-2.9)^2 + (2.1-2.9)^2 + (2.1-2.9)^2 + (7.1-2.9)^2}{4}} = 2.4$$

b) Compare the mean and median. Which is more resistant to outliers? (2 points)

The mean is higher than the median because it is heavily influenced by the single large value in the sample. The median is more resistant to outliers.

2) In the “March Madness” tournament of basketball there are 64 teams that begin the tournament. After each game one team is eliminated.

a) If the outcome of the tournament were completely random, what would be the probability of being in the final four (6 points)?

There are 4 teams in the final four, and there are 64 teams to start.

$$\boldsymbol{p = \frac{\textbf{number of teams in final four}}{\textbf{number of teams in tournament}} = \frac{4}{64} = \frac{1}{16} = 0.0625}$$

b) What is the probability of NOT being in the final four (4 points)?

This questions asks us to find the complement of $\boldsymbol{p}$ from part a.

$$\boldsymbol{p(\textbf{not in final four}) = 1 - p(\textbf{in final four}) = 1 - 0.0625 = 0.9375}$$

3) Joe wants to do a study to see whether having tattoos is related to having Hep C among people at San Quentin. To figure this out, he decides to test everyone coming through reception during the month of March for Hep C and also identifies whether these people have tattoos.

Joe finds no association between being positive for Hep C and having a tattoo. Based on this finding, Joe concludes that among people at San Quentin, <u>getting</u> a tattoo doesn't make someone more likely to get Hep C.

a) What type of study design is this (4 points)?

Observational

b) Joe makes a causal claim. Why doesn't his study support this? (6 points)

This was not a study to determine association not causation. So, stating that getting a tattoo does put someone at higher risk of acquiring Hep C cannot be a conclusion of this study. In addition, Joe did not take a sample of people at San Quentin. He took a sample of people in reception so he can't make a conclusion about people living at San Quentin based on this data.

4) Explain the following:

a) What is the placebo effect? Why is it important to randomize in clinical studies? (3 points)

The placebo effect is when an individual appears to benefit from treatment when no treatment is actually given. A patient that believes he or she is being treated may improve despite receiving no treatment at all. If the patients are not randomized, then patients will know they are receiving the treatment and may improve due to the placebo effect rather than the actual treatment. Randomization prevents the placebo effect being attributed to the treatment.

b) What happens to the standard deviation of the sample mean as the sample size gets larger (2 points)?

The sample mean ($\bar{x}$) is the mean of a sample of the population. Its standard deviation is the population standard deviation divided by the square root of the sample size. Therefore it gets smaller with larger sample sizes.

5) In past studies, it has been found that the distribution of weights for a certain population of young men ages 21 to 29 is normal with mean 180 pounds and standard deviation 25 pounds.

a. Given these facts, what proportion of this population weighs more than 200 pounds (5 points)?

We need to calculate the z-score so we can look the value up in the table.

$$z = \frac{X - \mu_x}{\sigma_x} = \frac{200 - 180}{25} = \frac{20}{25} = 0.8$$

The table shows the probability for z<0.8, so we need the complement of this.

$$p(z < 0.8) = 0.7881$$
$$p(z > 0.8) = 1 - P(z < 0.8) = 0.2119$$

Alternative: $p(z > 0.8) = p(z < -0.8) = 0.2119$

b. You are planning to take a simple random sample of 100 men from this population. What is the shape of the distribution of $\bar{x}$? What is the mean of $\bar{x}$? What is the standard deviation of $\bar{x}$ (5 points)?

$\overline{x}$ is the mean of a sample of a normally distributed population, so $\overline{x}$ will also be normally distributed.
The mean of $\overline{x}$ will be μ_x or 180.
The standard deviation of $\overline{x}$ will be $\sigma_x/\sqrt{n}$ or $25/\sqrt{100}$ which is 2.5

c. What is the probability that $\bar{x}$ is greater than 200 pounds (5 points)?

This is a z-score question again.

$$z = \frac{200 - 180}{2.5} = \frac{20}{2.5} = 8$$

We can't look this up in our table because the largest z value is 3.49. The probability of getting z>3.49 is less than 0.0002, and our z is much larger than this. We expect the probability of getting $\overline{x}$ greater than 200 pounds to be extremely small.

d. **[EXTRA CREDIT]** If your sample $\bar{x}$ turns out to be 210 pounds, what would you conclude about the population mean? Do you think the population mean could really be 180 pounds (5 points)?

It seems very unlikely that $\overline{x}$ could be 210 pounds with this population mean and standard deviation. If the standard deviation where much, much larger, that might make this possible, but the estimates for the population mean and standard deviation appear to be incorrect.

6) In city Y the average hourly income for a waiter is $23.90 with a standard deviation of $4.30.
   In city Z the average hourly income for a waiter is $21.50 with a standard deviation of $2.20.
   George waits tables in city Y and makes $25.30 per hour.
   Sam waits tables in city Z and makes $22.50 per hour.

   Assume the distribution of income is a normally distributed.
   Which waiter is making a higher income relative to the average in their city (10 points)?

   Hint: How do you compare the values of two random variables with different means and standard deviations?

To do this comparison, the z-score is our available tool. By using this technique, we are able to compare across different population means and standard deviations.

$$\boldsymbol{z = \frac{X - \mu_X}{\sigma_X}}$$

$$z(\text{George}) = \frac{\mathbf{25.30 - 23.90}}{\mathbf{4.30}} = \mathbf{0.326}$$

$$z(\text{Sam}) = \frac{\mathbf{22.50 - 21.50}}{\mathbf{2.20}} = \mathbf{0.45}$$

Sam is doing better than George relative to other waiters in their own cities.

7) There are three main shipping ports in California where large ocean going vessels load and unload shipping containers: Oakland, Long Beach and Los Angeles. These shipyards are managed independently, have different clients and experienced no major disruptions in the year of interest. Therefore, it is reasonable to assume that container movements at the ports are independent.

	Port of Los Angeles	Port of Long Beach	Port of Oakland
Mean number of containers loaded or unloaded per day in 2008	21,700	20,100	6,400
Standard deviation of containers loaded or unloaded per day in 2008	6,500	4,000	900

Find the mean and standard deviation of the number of shipping containers loaded or unloaded at California's three major ports each day in 2008 (10 points).

We have three variables to add. If we define $\boldsymbol{C}$ to represent the number of shipping containers loaded or unloaded, then we have the following:

$$\boldsymbol{C_{CA} = C_{LA} + C_{LBC} + C_{OAK}}$$

The means simply add in this linear combination of variables.

$$\boldsymbol{\mu_{CA} = \mu_{LA} + \mu_{LBC} + \mu_{OAK} = 21,700 + 20,100 + 6,400 = 48,200}$$

The standard deviations do not add linearly, but the variances do.

$$\boldsymbol{\sigma_{CA}^2 = \sigma_{LA}^2 + \sigma_{LBC}^2 + \sigma_{OAK}^2 = 6500^2 + 4000^2 + 900^2 = 59,060,000}$$

$$\boldsymbol{\sigma_{CA} = \sqrt{\sigma_{CA}^2} = \sqrt{59,060,000} = 7685 \approx 7700}$$

The daily total container loading and unloading at the three major ports in CA has a mean of 48,200 containers and a standard deviation of 7,700 containers.